主题与区域视角下我国城市政府开放数据利用现状分析*

■ 段尧清 邱雪婷 何思奇
华中师范大学信息管理学院 武汉 430079

摘要：[目的/意义]从数据层面分析我国城市政府开放数据在主题视角和区域视角下的利用现状，探究开放数据关注程度与利用程度之间的线性关系，以提高我国政府开放数据的使用效率。[方法/过程]以哈尔滨、济南、上海、武汉、广州和贵阳 6 地为例，对开放数据的浏览率、下载率等多个指标进行统计比较分析、聚类分析和回归分析，揭示我国城市开放数据在主题与区域视角下的利用现状，并探讨浏览率与下载率之间的线性关系。[结果/结论]我国城市政府开放数据的利用具有以下特征：在整体上开放数据浏览率与下载率呈弱相关。在主题视角下，教育科技、民生服务、经济工商等与社会民生领域相关的主题利用程度高，其浏览率与下载率呈正相关；开放数据的整体特征与部分特征具有不一致性。在区域视角下，济南、上海的浏览率和下载率呈正相关，其中上海对开放数据的利用排名首位，济南则排在末位，贵阳开放数据浏览率与下载率都较高，但二者呈弱相关；开放数据整体特征与部分特征大体上具有一致性。

关键词：政府开放数据 相关性分析 利用
分类号：D63 G203
DOI：10.13266/j.issn.0252-3116.2018.20.008

1 引言

政府开放数据是指由政府、政府委托和控制的实体产生的能被任何人自由地利用，再利用和再分配的数据[1]。该运动最早于 2009 年在美国兴起，国内外政府和学术界对其高度重视。据 2016 年 4 月发布的“开放数据晴雨表”全球报告(Open Data Barometer)显示，全球已有 114 个国家加入了这一行列[2]。我国以 2012 年“上海市政府数据服务网”的上线试运行为开放数据的标志，截止 2017 年 11 月，全国共有 23 个省、市或区政府建立了地方性政府数据开放平台(港澳台地区除外)。政府开放数据的利用是指已开放的数据资源满足人们需求和利用的情况与程度，其本质是资源的有效配置和使用。当前国内外对政府开放数据利用现状的研究比较宏观，并取得了一定的成果。

国外学界对政府开放数据利用的研究主要集中在两个方面：第一是开放数据的利用障碍研究。起初认为政府数据的开放水平对政府数据的服务效果有直接影响[3]，同时，开放数据中存在实施障碍和使用障碍[4]，随后又发现开放数据在商业利用中存在可访问性、可用性和交互性等障碍[5]，在此基础上构建评测模型识别用户的使用与满意度[6]，进而分析障碍机制。第二是开放数据的利用价值研究。在理论上，从评估各地政府开放数据的动机中[7]意识到政府开放数据的潜在价值应被有效的利用模式激活[8]，在实践中，相关概念模型[9]和可视化方法[10]被运用于挖掘、理解并传递开放数据的价值。

国内学界对开放数据利用的研究主要集中在 3 个方面：第一是对开放数据利用的评价研究。针对政府数据开放平台，提出相应的评估框架、指标和方法[11]，从用户利用的角度评估了时下我国已有的政府数据开放平台[12]，又根据服务绩效将广东和北京、上海等地的若干平台划分成了三个级别[13]；针对开放数据，主

* 本文系国家社会科学基金重点项目“基于全生命周期的政府开放数据整合利用机制与模式研究”(项目编号:17ATQ006)和中央高校基本科研业务费专项资金重大培育项目“大数据环境下的政府信息服务研究”(项目编号:CCNU16Z02002)研究成果之一。
作者简介：段尧清(ORCID:0000-0002-8991-5842)，教授，博士生导师；邱雪婷(ORCID:0000-0003-0842-5807)，硕士研究生；何思奇(ORCID:0000-0002-8186-6775)，硕士研究生，通讯作者，E-mail:Daisy_hsq@163.com。
收稿日期:2018-04-02 修回日期:2018-06-02 本文起止页码:65-76 本文责任编辑:杜杏叶

要选用数据集访问量、下载量、申请情况[14]、下载量与浏览量的比值，以及平均浏览量[15]等指标衡量政府开放数据的利用效果。第二是对开放数据的保障机制研究。政府开放数据的高效利用离不开法律、技术、数据共享和用户参与的充分保障[16]，同时开放数据的完整性、准确性、及时性和国家相关政策也会影响数据的利用[17]。第三是开放数据的利用方式研究。从宏观上探究影响政府开放数据利用的因素[18]，同时从微观上将 API 接口和 APP 程序开发等利用方式纳入数据利用范畴[19]，通过梳理国内外政府数据开放利用方式，提出了开放数据的价值的提高与数据的利用率成正比的观点[20]。

综上，国内外研究重在从宏观上探究数据开放平台的建设情况，鲜有从微观视角剖析开放数据的利用现状。政府开放数据的最终目的是促进其使用与开发，为了帮助政府最大限度满足公众数据需求，了解我国城市政府之间数据利用的差距，以数据本身作为切入点，分别从主题视角和区域视角出发，运用数据浏览率、下载率等指标对比分析开放数据的利用程度，通过聚类分析和相关性分析把握不同主题、不同城市政府开放数据的利用现状，促进我国政府开放数据工作的发展。

2 样本选取与数据采集

2.1 样本选取

基于我国政府数据开放平台大多数是以“data.gov.cn”为域名的，因此，本文以“data.gov.cn”为域名进行搜索，截止 2017 年 11 月 29 日，我国已有 23 个城市建立了开放数据平台，从现有平台来看，开放数据主要集中在经济、交通、教育、环保等社会生活领域的各方面。但是，各城市平台开放的数据主题在数量和名称上有较大差异，存在同类数据在不同平台名称不同，以及不同主题所含的数据同属一个大类的情况。因此，本文首先梳理现有 23 个数据开放平台的主题分类情况，为了提高研究的集中性和效率，将现有平台全部数据集的 297 个资源主题整理归纳为经济工商、财税金融等 19 个大类，并统计了各大类中子主题的分布情况，具体分布情况见图 1。

2.1.1 主题视角下研究样本的选取　从我国政府开放数据主题分类统计中可以看出，现有开放数据主题的分布呈集中与分散的状态。在开放数据的数量上，“文体休闲”类共包含 29 个子主题，是所含类目数量最多的主题大类，类目数量最少的是“宗教信

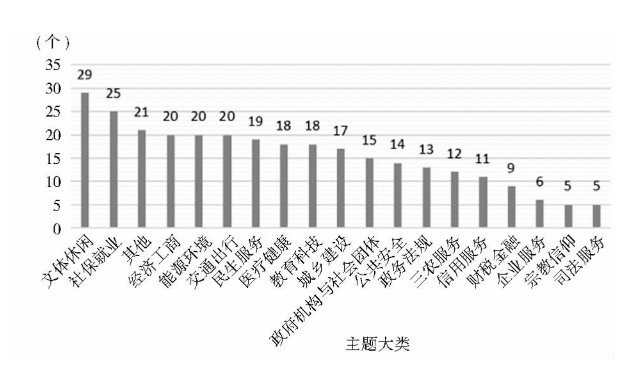


图 1　我国政府开放数据主题分类统计

仰”和“司法服务”大类，它们都分别包含 5 个子主题；在开放领域上，当前我国地方政府开放数据重点集中在与教育、就业、医疗等相关的民生领域（文体休闲、社保就业、交通出行、教育科技、医疗健康等）、生态领域（能源环境）和社会治理领域等（政府机构与社会团体、公共安全等）。主题分类表明各地方政府对开放数据的分类既具有一定的共性，又带有某些个性。

依据主题分类共性特征（指某些主题同时出现在平台上）和区域样本（区域样本的选取见 2.1.2）的分布特点及其在各平台的覆盖面，剔除“其他类”后，选取“文体休闲”“社保就业”“经济工商”“能源环境”“交通出行”“民生服务”“教育科技”“政府机构与社会团体”“公共安全”10 个主要大类作为主题视角的初步研究样本。

2.1.2 区域视角下研究样本的选取　开放数据的城市在行政级别和地域上具有广泛的代表性。在行政级别上，这 23 个城市包含省级城市、副省级城市、地级市、县级市等多种行政级别；在地域上，它们分散于北京、上海、湖北、广东等多个直辖市或省份。

由于行政级别存在差异，各开放平台开放数据的侧重点与进程差别较大，不适合做全样本分析。因此，经过多次讨论，本文拟选取济南、上海、武汉、广州、哈尔滨和贵阳 6 个城市为研究样本。这 6 个城市的可比性体现在两个方面：一是除了武汉在“社保就业”和贵阳在“民生服务”主题中尚未涉及以外，其余 8 个主题大类均能与 6 个研究样本相对应；二是这 6 个样本在行政级别上均为直辖市或省会城市，且它们在地理位置上均分属不同区域，因此这 6 个城市作为区域样本具有一定代表性和意义。本文结合主题视角和区域视角对研究样本的确定，选取 6 个平台中的 63 个子主题，将其归纳到 10 大主题大类中，最终确定的研究样本如表 1 所示：

表 1　我国省会城市与直辖市政府数据开放平台情况

平台 \ 主题大类	文体休闲	社保就业	经济工商	能源环境	交通出行	民生服务	医疗健康	教育科技	政府机构与社会团体	公共安全	总计
济南市政府数据开放平台	（文化、体育）（旅游、服务业）	扶贫救灾（劳动、人事）	经济管理	（城乡建设、环境保护）（国土资源、能源）	（工业、交通）	（民政、社区）	卫生健康	（科技、教育）	综合政务	（政法、监察）	13
上海市政府数据服务网	文化休闲	社会发展	经济建设	资源环境	道路交通	民生服务	卫生健康	教育科技	机构团体	公共安全	10
武汉市政务公开数据服务网	文化娱乐	缺	经济发展	能源环境	交通服务	公共服务	医疗卫生	教育科技	政府机构	公共安全	9
广州市政府数据开放平台	文体娱乐	社会发展 劳动人事	经济发展	资源环境	道路交通	民生服务	健康卫生	教育科技	机构团体	公共安全	11
哈尔滨市政府数据开放平台	文体休闲	社会发展	经济建设	资源环境	道路交通	民生服务	卫生健康	教育科技	机构团体	公共安全	10
贵阳市政府数据开放平台	文体休闲	社会发展 劳动人事	经济建设	生态文明	交通运输	缺	卫生健康	教育科技	政府机构	公共安全	10
总计	7	8	6	7	6	5	6	6	6	6	63

2.2　数据采集

据统计，以上 6 个政府数据开放平台一共有 86 个子主题，本文研究涉及的子主题有 63 个，利用八爪鱼数据采集器采集与人工观察的方式分平台抓取以上 63 个子主题数据集的相关信息，以供后续实验使用。所有数据的采集截止日期为 2017 年 12 月 12 日。

2.3　开放数据利用现状的参数确定

当前研究大多从浏览量和下载量等宏观角度衡量开放数据的利用效果，学者们认为数据的浏览量和下载量会影响用户对数据的关注和利用效果[21]，并通过计算下载量与浏览量的比值来比较北京和上海开放数据的利用效果[22]，有的还选取访问量和下载量来衡量时下我国部分地方政府数据的利用状况[13]。但是，浏览率和下载率等指标更能反映局部数据在整体数据中的占比情况，进而反映用户对某主题或某区域开放数据的关注程度和利用程度，因此本研究选取开放数据的浏览率、下载率等指标作为测算参数，具体参数设计如下。

设表 1 中主题视角下的 10 大类样本全部主题的开放数据集的总量为 NT，浏览总量为 BT，下载总量为 DT。具体公式如(1)－(3)所示：

$$NT = \sum_{i=1}^{10} nt_i \quad (1)$$

$$BT = \sum_{i=1}^{10} bt_i \quad (2)$$

$$DT = \sum_{i=1}^{10} dt_i \quad (3)$$

其中，每一个大类主题都包括若干个子主题。i 代表 10 个主题大类(i＝1，2，…，10)；nt_i 表示第 i 主题大类下开放数据集的数量；bt_i 代表第 i 主题大类开放数据的浏览量；dt_i 代表第 i 主题大类开放数据的下载量。

同理，设表 1 中 6 个城市全部主题的开放数据集总量为 NC，浏览总量为 BC，下载总量为 DC。具体公式如(4)－(6)所示：

$$NC = \sum_{i=1}^{6} nc_i \quad (4)$$

$$BC = \sum_{i=1}^{6} bc_i \quad (5)$$

$$DC = \sum_{i=1}^{6} dc_i \quad (6)$$

其中，i 代表 6 个城市(i＝1，2，…，6)，nc_i 表示第 i 城市下开放数据集的数量，bc_i 代表第 i 城市开放数据的浏览量，dc_i 代表第 i 城市开放数据的下载量。

在此基础上，首先统计分析主题和区域视角下开放数据的浏览率和下载率，以揭示用户对开放数据的关注情况和利用程度，再对浏览率与下载率做回归分析，以找出二者之间的相关性，帮助实现政府数据的价值和目的。依据上述公式，其他测算公式及方法见表 2。

3　数据分析

3.1　主题视角下政府开放数据的利用现状

为了揭示不同主题视角下开放数据的利用现状，首先统计各主题数据的浏览量(bt_i)、下载量(dt_i)、单一样本开放数据平均浏览率(bt_i/nt_i)和平均下载率(dt_i/nt_i)，如图 2、图 3 所示，同时结合开放数据浏览率(bt_i/BT)和下载率(dt_i/DT)的对比分析来比较各主题数据的利用现状。

3.1.1　主题视角下开放数据浏览率　浏览率能直观反映用户对某一主题数据的关注情况，首先运用 R 绘制各主题大类开放数据的浏览率折线图和浏览率散点图，如图 4 所示。其中 X 轴代表 10 个主题大类，也即 t_i，(i＝1，2，…，10)。主题 1－10 分别代表：文体休闲、经济工商、交通出行、医疗健康、政府机构与社会团体、社保就业、能源环境、民生服务、教育科技和公共安全。

表 2　开放数据利用现状测算指标参数及方法

测算指标	开放数据浏览率	开放数据下载率	单一样本开放数据平均浏览率	单一样本开放数据平均下载率	整体样本开放数据平均浏览率	整体样本开放数据平均下载率
主题视角	$\frac{bt_i}{BT}$	$\frac{dt_i}{DT}$	$\frac{bt_i}{nt_i}$	$\frac{dt_i}{nt_i}$	$\frac{bt_i}{nt_i}/\sum_{i=1}^{10}\frac{bt_i}{nt_i}$	$\frac{dt_i}{nt_i}/\sum_{i=1}^{10}\frac{dt_i}{nt_i}$
区域视角	$\frac{bc_i}{BC}$	$\frac{dc_i}{DC}$	$\frac{bc_i}{nc_i}$	$\frac{dc_i}{nc_i}$	$\frac{bc_i}{nc_i}/\sum_{i=1}^{6}\frac{bc_i}{nc_i}$	$\frac{dc_i}{nc_i}/\sum_{i=1}^{6}\frac{dc_i}{nc_i}$

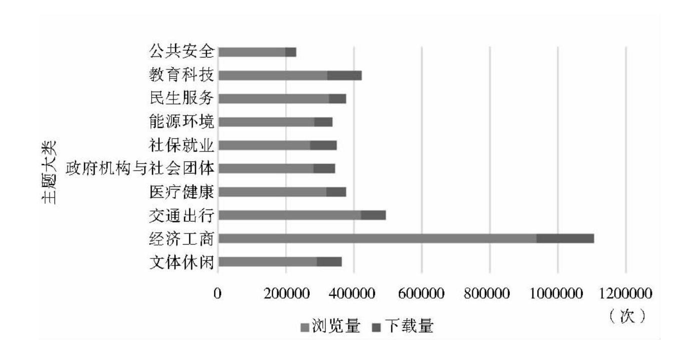


图 2　主题视角下开放数据的浏览量和下载量

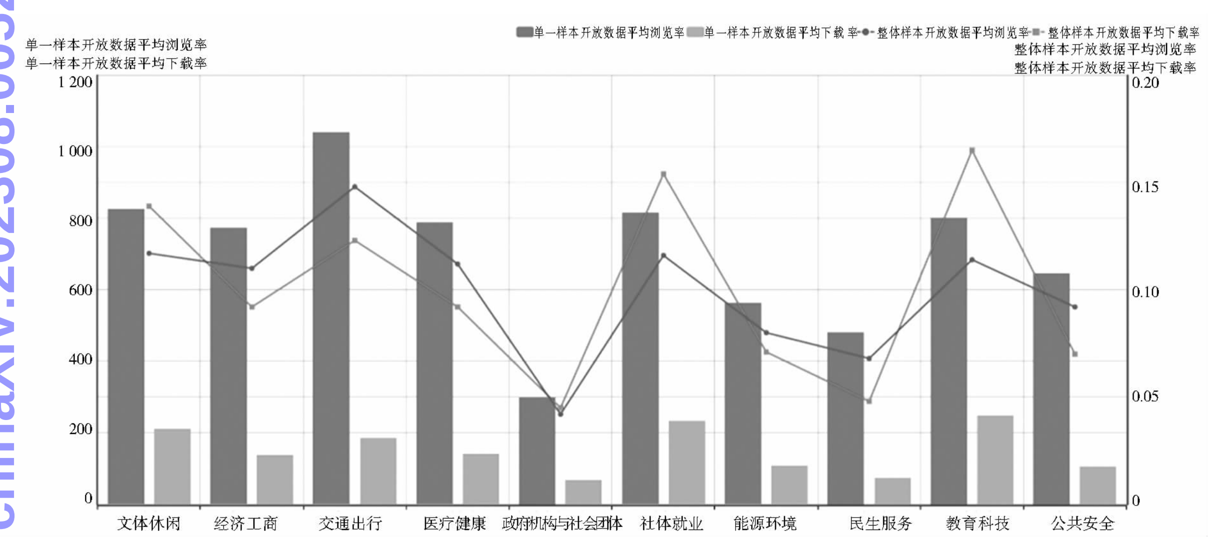


图 3　主题视角下单一(整体)样本开放数据的平均浏览(下载)率

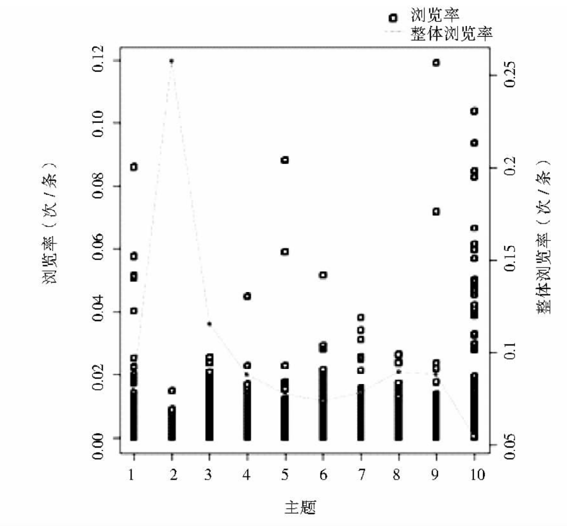


图 4　主题视角下开放数据浏览率折线图和散点图

从图 2 可知，浏览量排在前三位的是经济工商(937 657 次)、交通出行(420 002 次)和民生服务(327 778次)，其中交通出行类数据的单一样本平均浏览率也是最高的，由此可知，经济工商和交通出行等主题的数据备受用户关注；然而结合图 3 可知，仍然有部分主题在该指标的值明显低于平均值，如政府机构与社会团体主题数据的平均浏览率均不足 300 次/条，这表明各主题数据被关注的程度存在差异性。而图 4 则显示浏览率最高的主题经济工商(约为 0.267)是最低的公共安全(约为0.054)的4.94 倍，且这 10 类数据中仅有经济工商和交通出行两类数据的浏览率高于平均值(0.1)。

此外，对比图 4 中的开放数据的浏览率折线图和

散点图可知，经济工商类数据的浏览率虽然最高，但由其散点图的分布可知，并非其开放任意一条数据的浏览率都很高(大多数落在0-0.02以内)。而公共安全类的数据却恰好相反，虽然该数据集的整体浏览率最低，但其开放数据浏览率的跨度较大，大多在0.1-0.25之间，这表明各主题开放数据整体与部分的特征并不具有一致性。

3.1.2 主题视角下开放数据下载率 下载率是对浏览率的进一步说明与深化，它在很大程度上能反映出用户对某一数据的利用情况。同浏览率一样，图5所示的是各主题大类开放数据的浏览率折线图和散点图，主题1-10的含义与3.1.2中图4相同。

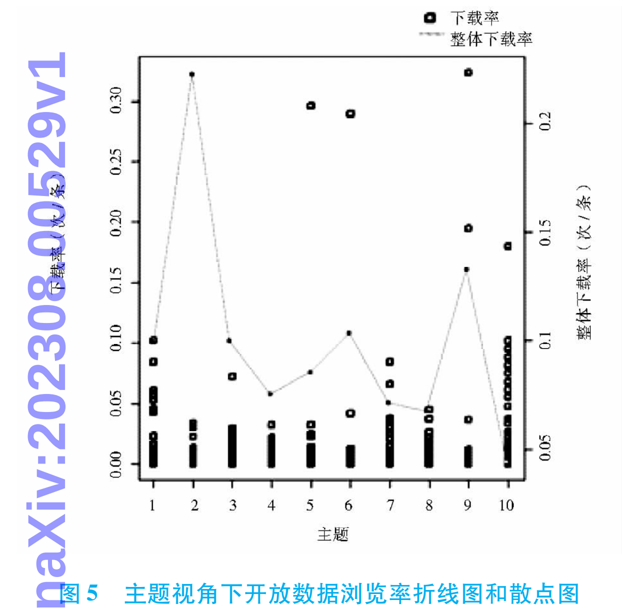


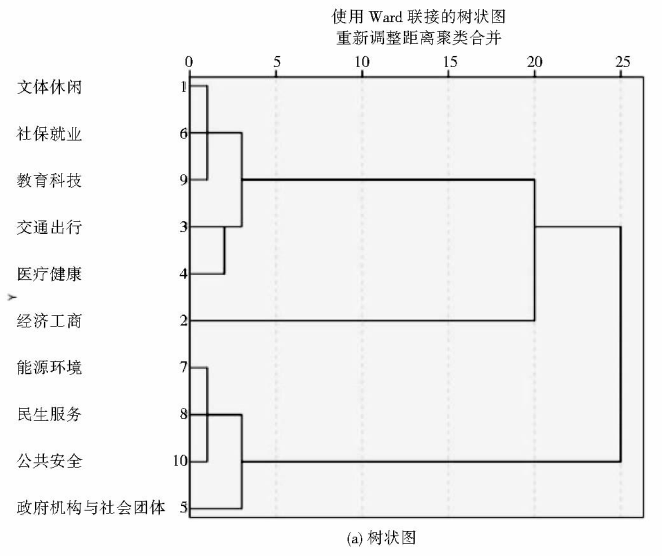
图5 主题视角下开放数据浏览率折线图和散点图

由图5可知，用户对经济工商、教育科技、社保就业和交通出行类数据的利用程度较深，体现在2个方面。第一，开放数据下载率高于平均值的有4个主题，也即经济工商(约0.223)、教育科技(约0.133)、社保就业(0.103)和交通出行(约0.1)，同时，除去社保就业以外，其余3个主题的整体开放数据平均浏览率也均高于相应平均值，这间接表现出用户对此类数据利用的真实性；第二，经济工商等4个主题数据的下载率之和约为0.56，占所有主题开放数据下载总量的一半以上，表明它们被利用的程度较高，同时也反映出其余几类数据的利用程度有待提高。

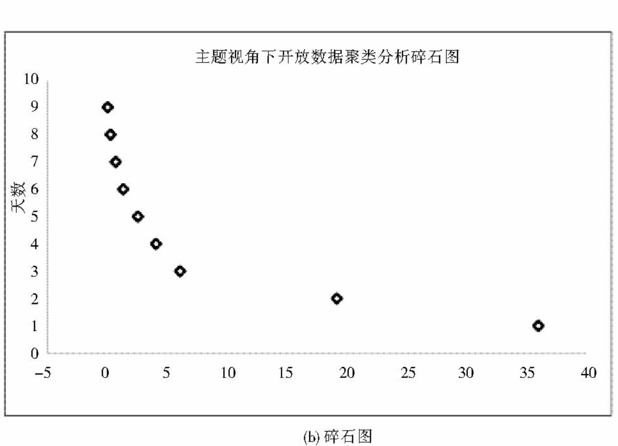
此外，结合下载率散点图可知，其余各主题下载率分布不均衡，首先，以公共安全为例，其下载率散点图的跨度虽然较大，但总体上该主题的下载率排在末尾；其次，包括经济工商在内的大部分主题整体开放数据的平均浏览率都不高，其值大多落在0-0.05以内。

3.1.3 主题视角下开放数据的利用现状 为了深入揭示各主题开放数据的利用现状与亲疏关系，对10个主题大类做聚类分析。

聚类分析是指在事先不规定分组规则的情况下，将数据按其自身特征划分成不同的群组，各群组内部数据差距尽可能的小，而各群组数据之间的差距尽可能的大[23]。首先选择聚类指标，由于单一样本和整体样本的开放数据平均浏览率(下载率)的变化方向一致且一一对应，前者是测算后者的基础，因此聚类分析主要参考开放数据浏览率、下载率和整体样本开放数据平均浏览率(下载率)4个指标的相关情况。与此同时，选用层次聚类法，其中个体距离采用平方欧式距离，类间距离采用Ward联接，最终聚类结果如图6所示：



(a) 树状图



(b) 碎石图

图6 主题视角下开放数据聚类分析的树状图和碎石图

为了更好地划分聚类数目，绘制了主题视角下开放数据聚类的碎石图，如图6(b)所示，随着类的不断凝聚和类目数量的不断减少，各类之间的距离迅速增大，碎石图逐渐趋于平坦。观察碎石图可知，当聚成4类之前，各类之间的距离较小，当聚成4类之后，各类之间的距离较大，由此可知，4类就是该碎石图的"拐点"，因此聚成4类或3类较好。经过综合考虑，本研究最终将10大主题聚为4类，见图6(a)。

具体而言，第一类是"文体休闲""社保就业"和"教育科技"，仅经过三步就聚成一类，其系数分别为0.128和1.465。社保就业和教育科技类的数据在整体样本开放数据平均浏览率等3项指标均超过相应平均值，其余各项指标均排名较前，方差与标准差都较小。这表明与民生相关的数据利用效率较高，教育科技、社保就业和文体休闲是用户最为关心的日常问题之一，与用户的距离最近，也进一步说明用户需求是数据利用的前提。

第二类是"经济工商"。经济工商在浏览率、下载率等4项指标中占据绝对优势，因此自成一类，该类数据较高的关注度与利用率主要是由上海市政府的开放数据贡献的，此类数据主要涉及某地经济建设和工商贸易等信息，涵盖了经济、工商、统计、贸易、消费、经济政策信息等方面，用户尤其是企业用户对此类数据的关注和需求更大，因此经济工商类数据各项指标都稳居高位。

第三类是"交通出行"和"医疗健康"类数据，它们在第5步时与医疗健康聚成一类，二者之间的系数为5.187，它们虽然在总体上浏览率与下载率不高，但其整体样本开放数据平均浏览率与下载率却排在前列。交通出行与用户的生活联系紧密，而医疗健康更是全社会关注的热点，当下"互联网+"交通和电子医疗的出现，大大节约了用户的时间，便利了公众的生活。

第四类是"民生服务""能源环境""政府机构与社会团体"和"公共安全"类数据，它们的4项指标均为负，且低于平均值。这一类数据的利用相对不高，用户对能源环境、公共安全等社会治理领域的问题目前关注还不太多，与用户意识、需求的紧急性等因素有关。

3.2 区域视角下政府开放数据利用现状

本节从浏览率(bc_i/BC)、下载率(dc_i/DC)及其对比分析三方面出发，首先统计了各地区的浏览量(bc_i)和下载量(dc_i)，如图7所示；同时绘制了单一样本开放数据平均浏览率(bc_i/nc_i)和下载率(dc_i/nc_i)及整体样本开放数据平均浏览率($\frac{bc_i}{nc_i}/\sum_{i=1}^{6}\frac{bc_i}{nc_i}$)及下载率($\frac{dc_i}{nc_i}/\sum_{i=1}^{6}\frac{dc_i}{nc_i}$)统计图，见图8。

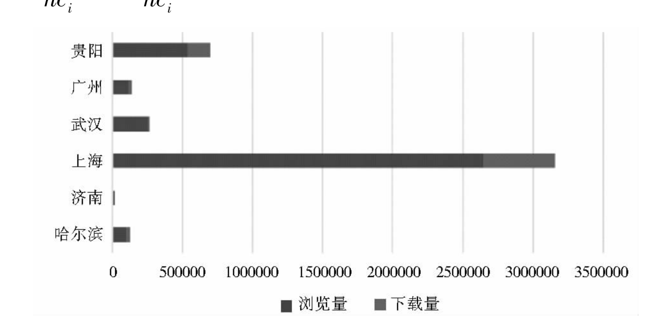
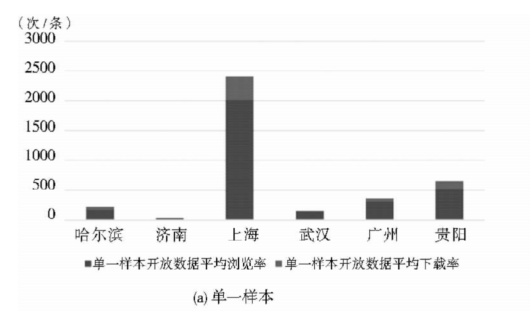


图7 各地区整浏览量和下载量



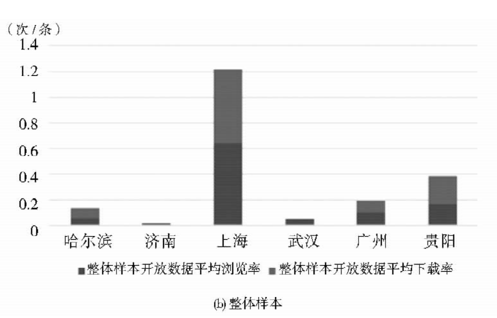


图8 各地区单一(整体)样本开放数据平均浏览率和下载率

3.2.1 区域视角下开放数据浏览率 同主题视角下开放数据的浏览率一样，将各区域的浏览率折线图与散点图绘制在一起，如图9所示，便于观察各地区开放数据的关注程度与状况。其中，X轴代表6个城市，也即c_i，(i=1,2,…,6)。

结合图8(b)和图9可知，不论是各区域浏览率还是其整体样本的平均浏览率，上海和贵阳都是表现最好的城市。以上海为例，其数据浏览率是济南的166.883倍，同时其整体样本的平均浏览率也达到0.63，远高于平均值(0.1667)；而济南正好相反，这两项指标均排在末尾，其余各地排名稍有变化。此外，上海和贵阳两地开放的交通出行主题数据的浏览率之和

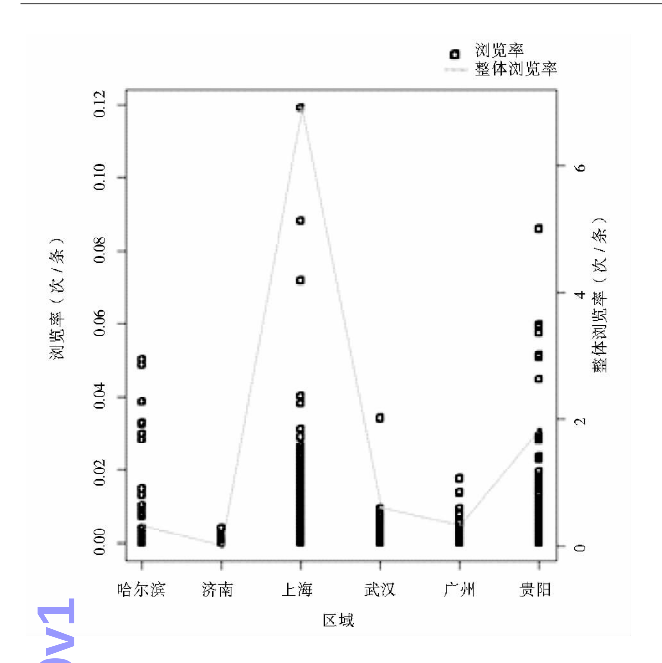


图 9　区域视角下开放数据浏览率折线图和散点图

占 6 地交通出行类数据总浏览率的 89.3%。

此外，结合浏览率散点图可知，上海和贵阳不仅在整体上开放数据浏览率高，其散点图的跨度也较大，其中，浏览率落在平均值(1.67)以上的数据量较多；除了上海和贵阳以外，其余城市浏览率排名依次是武汉、广州、哈尔滨、济南，其中，哈尔滨虽然整体浏览率不高，但相应的散点图跨度较大，且其政府数据开放平台上存在相当数量的浏览率高于平均值的数据。

3.2.2　区域视角下开放数据的下载率　分析各区域开放数据的下载率有利于分析用户对各城市开放数据的利用程度。各区域下载率的折线图与散点图见图 10。

由图 8 和图 10 可知，上海和贵阳分别作为一个整体，其下载率较高，同时其样本的平均下载率也占据了较大优势，例如上海作为开放数据下载率最高的城市，

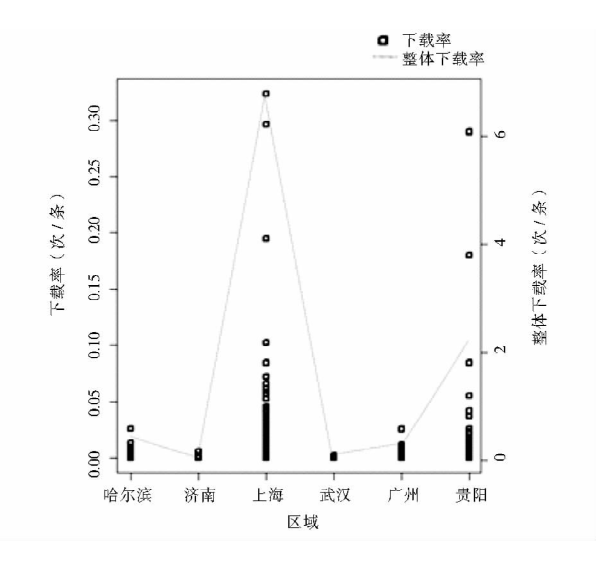


图 10　区域视角下开放数据浏览率折线图和散点图

它与最低的济南相差 136 倍之多；此外，上海除社保就业外的其他 9 个主题数据均拥有最高下载率，且全都超过平均值(约 0.167)，其中教育科技的下载率更是高达 0.798。但同时也有城市的开放数据下载率较低，如武汉的开放数据在各项下载率的指标上均排名末尾。

结合各城市开放数据的散点图可知，同各城市开放数据浏览率一样，下载率高的城市，其散点图的跨度也更大；反之，整体下载率低的城市，其下载率跨度相对来说较小，但这并不代表个各城市开放数据浏览率与下载率的变化呈正比。

3.3.3　区域视角下开放数据的利用现状　为了进一步揭示政府开放数据在区域视角下的利用现状，结合浏览率等 4 个指标，采用层次聚类法对各区域开放数据的利用状况做聚类分析。其中，个体距离采用平方欧式距离，类间距离采用平均组间联接，最终聚类结果的冰柱图和树状图如图 11 所示：

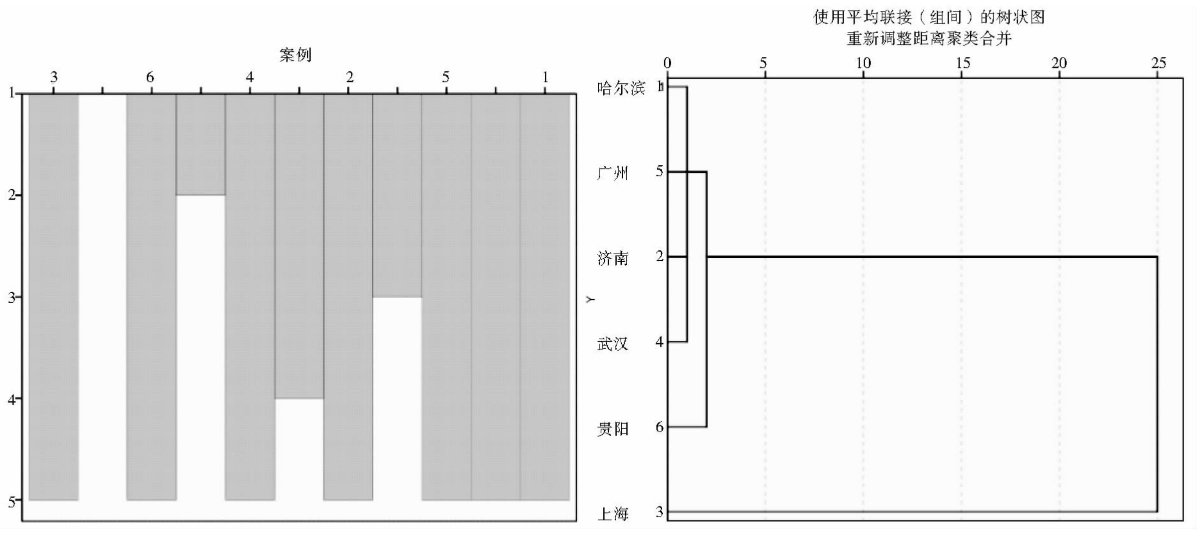


图 11　区域视角下开放数据聚类分析冰柱图和树状图

由于区域研究样本较少,因此直接采用观察法对其聚类数目进行划分。据冰柱图可知,当聚成4类时,哈尔滨和广州为一类、济南和武汉为一类、贵阳和上海分别单独聚成一类;当聚成3类时,哈尔滨、广州、济南和武汉为一类、贵阳和上海分别单独为一类。为了更为细致的分析开放数据利用的分布研究,将6个研究样本聚成4类。

第一类是哈尔滨和广州。两地具有较高的相似性,首先聚为一类,系数仅为0.043。在4个衡量指标上,哈尔滨和广州的指标排名相差不大,紧跟在上海和贵阳之后,各项指标大多处于中间位置。同时,哈尔滨和广州开放数据的相似性还体现在其数据开放的起始时间、开放数据的数量、格式等方面,虽然起步较晚,但其政府数据的关注度和被利用情况尚可。

第二类是济南和武汉。这两个城市之间的系数为0.125,相似度较高。武汉虽然比济南早2年开放数据,开放数据集总量大,有较高的浏览率,但由于受到开放格式等因素的影响,武汉开放数据的整体和平均下载率都很低;同样,济南开放数据起步晚,数据集量少,在4项指标中的排名都十分靠后。因此与武汉的差距并不明显,这两地聚为一类。

第三类是贵阳。贵阳虽然开放数据起步较晚,但在不足一年的时间内就取得数据开放指数排名第二的成绩[24]。除整体样本的平均浏览率外,贵阳开放数据的其余各项指标均超过平均值,此外,其开放的9个主题数据集(民生服务主题暂缺)的浏览量均稳居前三,贵阳开放数据的关注程度和利用程度仅次于上海。

第四类是上海。上海开放数据的浏览率和下载率均在首位,自从2012年开放数据以来,上海通过政府引导、提高数据质量、重视用户参与和数据创新等方式,使得其在各地方政府开放数据中稳居第一梯队[25];在本研究中,上海的开放数据在浏览率、下载率等4个指标中均排名第一,其中经济工商类数据的单一平均下载率高达352次/条(约数),是其他类数据的2倍之多。

4 浏览率和下载率的相关性分析

4.1 开放数据整体的浏览率与下载率的相关性分析

浏览率与下载率分别代表用户关注程度与数据利用程度,探明二者之间的关系有利于帮助提高数据的利用效率。首先在不区分研究视角的条件下用R绘制了浏览率与下载率的散点图,见图12。

如图12所示,代表浏览率与下载率的散点大量的

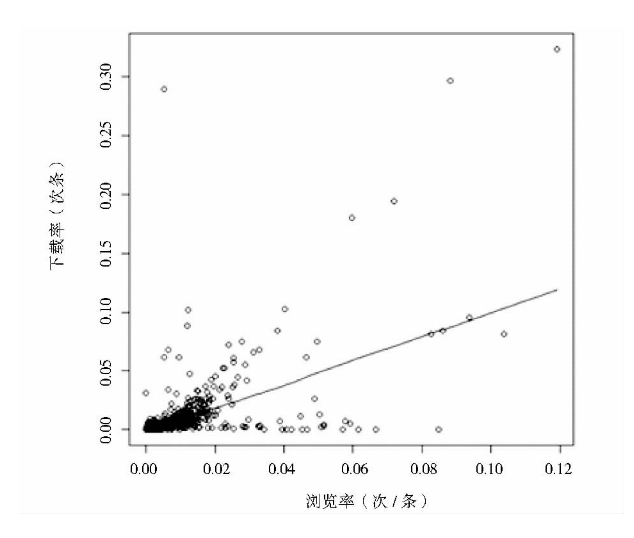


图12 浏览率与下载率的散点图

分布在趋势线两侧,仅有部分数据呈现出浏览率越高、下载率越高的趋势。为了进一步分析浏览率与下载率之间的关系,运用R回归模型拟合上述数据,首先运用线性最小二乘法做回归分析,回归结果显示回归系数p值(0.166, <2e-16)很小,非常显著的≠0; *** 也表示显著程度非常显著。同时F统计量=3799,p-value: < 2.2e-16远小于0.05,表示整个回归模型显著,适合估计download_rate变量。拟合优度 R^2 = 0.4058 < 0.5,表示拟合程度较弱,在此基础上绘制回归诊断图。

图13中的4张诊断图分别是(1)残差与拟合值图,图中无明显曲线关系;(2)残差Q-Q图,说明实验数据不服从正态分布;(3)标准化残差与拟合值图,纵坐标是标准化残差的平方根,残差越大,点的位置越高,模型残差等方差;(4)残差与杠杆图,鉴别出了离群点、高杠杆点、强影响点。

综上所述,浏览率与下载率在整体上虽然存在一定相关性,但相关程度较弱。

4.2 主题视角下开放数据浏览率与下载率的相关性分析

开放数据浏览率与下载率在整体上存在相关性,但并不能说明各主题开放数据浏览率与下载率的关联程度。因此,分主题分别对10个主题浏览率与下载率做回归分析得到表3,同时绘制各主题开放数据浏览率与下载率的散点图,如图14所示,以分析其在不同主题下的变化规律。

结合表3和图14, R^2 值越接近于1,表明浏览率和下载率相关性越强,也即开放数据的浏览率越高,下载率也越高,同时也表明数据关注程度与利用程度关联性越大。

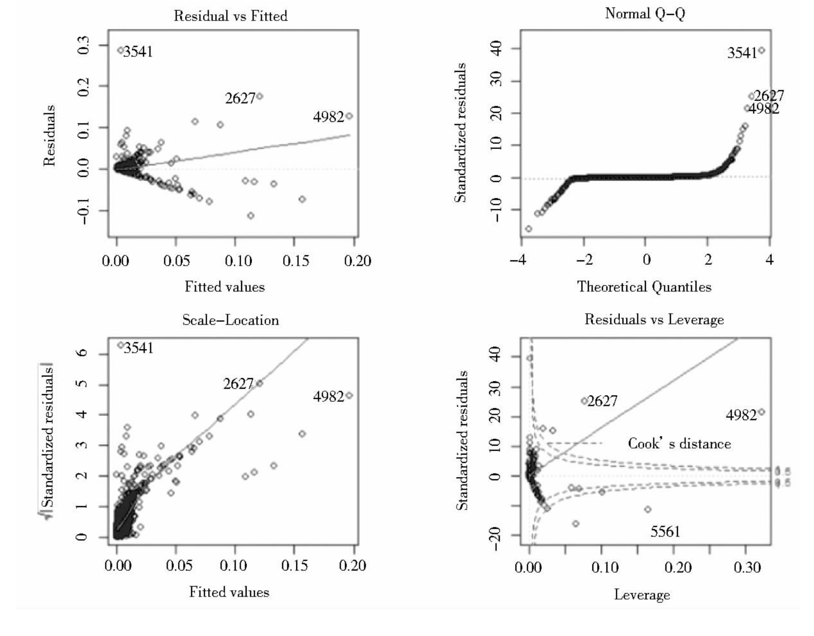


图 13　回归诊断图




表 3　主题视角下开放数据浏览率与下载率的相关性分析

城市	R^2	主题	R^2
教育科技	0.863 3	交通出行	0.594 6
民生服务	0.830 5	文体休闲	0.505 6
经济工商	0.704 4	公共安全	0.304 9
医疗健康	0.629	社保就业	0.015 28
政府机构与社会团体	0.611 2	能源环境	0.015 28

在主题视角下，共有 8 个大类的浏览率与下载率呈现出不同程度的正相关关系。首先，教育科技和民生服务主题的浏览率和下载率呈现较强的正线性关系，其 R^2 系数分别表示回归关系能解释因变量 86.33% 和 83.05% 的变异，回归效果较好，这其中，教育科技类数据的浏览率与下载率基本成正比，且其散点图分布较为均匀；其次，经济工商的回归相关系数为 0.704 4；此外，社保就业和能源环境的 R^2 分别均为 0.015 28，表明这两个主题的数据浏览率与下载率不相关。

因此，主题视角下浏览率与下载率关联程度由强到弱：教育科技 > 民生服务 > 经济工商 > 医疗健康 > 政府机构与社会团体 > 交通出行 > 文体休闲 > 公共安全 > 社保就业 = 能源环境。

4.3　区域视角下开放数据浏览率与下载率的相关性分析

为了统计各区域开放数据浏览率与下载率的关联程度，分区域对 6 个城市开放数据浏览率与下载率做回归分析，分析结果如表 4 所示，与此同时绘制 6 个城市开放数据的散点图，如图 15 所示。

表 4　区域视角下浏览率与下载率相关性分析

城市	R^2	城市	R^2
济南	0.780 3	哈尔滨	0.443 1
上海	0.753 4	武汉	0.318 4
广州	0.625 1	贵阳	0.153 7

由表 4 可知，区域视角下 6 个城市开放数据的浏览率与下载率呈正相关，但其相关性呈现出一定的差异性。

具体来说，济南开放数据的浏览率和下载率呈现较强的正线性关系，R^2 系数为0.780 3，这在区域浏览率与下载率散点图中也可以得到验证；上海紧跟其后，其数据的下载率与浏览率的回归相关系数为0.753 4；此外，武汉和贵阳相关系数的平方分别为0.318 3、0.153 7，表示回归关系仅能解释因变量 31.83%、15.37% 的变异，回归效果较差，因此武汉和贵阳开放数据的浏览率与下载率的关联程度最弱。总之，区域视角下浏览率与下载率关联程度由强到弱为：济南 > 上海 > 广州 > 哈尔滨 > 武汉 > 贵阳。

5　结论与讨论

（1）通过聚类分析，得出以下结论。第一，从主题视角看，用户对经济、民生等与日常生活联系密切的领域关注更高，10 个不同主题数据的利用现状呈现出一定的差异性，将 10 个大类的数据根据相似性划分成了 4 类：文体休闲、社保就业和教育科技是第一类；经济工商单独成第二类；交通出行和医疗健康是第三类；民生服务、能源环境、政府机构与社会团体和公共安全聚

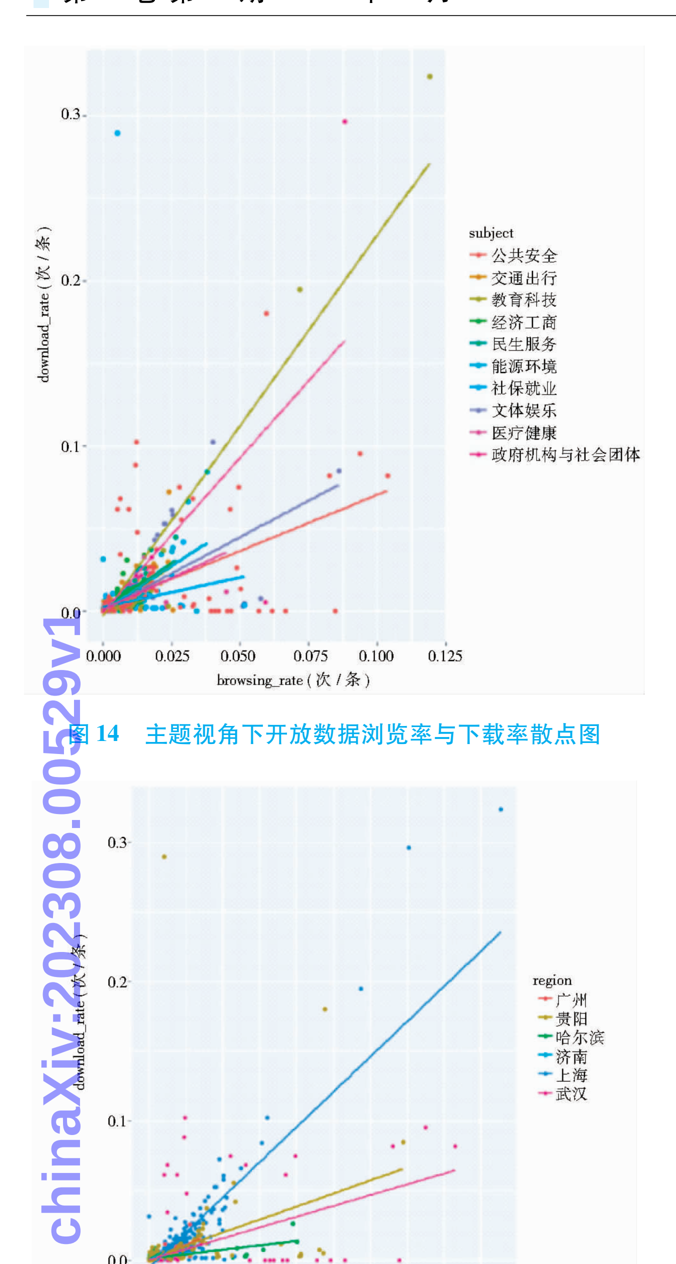


图 14 主题视角下开放数据浏览率与下载率散点图

图 15 区域视角下开放数据浏览率与下载率散点图

成第四类。第二，从区域视角看，不同城市开放数据的利用程度不同，呈现出明显的高低之分：哈尔滨和广州为第一类、济南和武汉是第二类、贵阳和上海分别单独聚成第三类和第四类。

（2）通过回归分析发现开放数据浏览率与下载率之间的变化规律，用以探究开放数据利用情况。

①首先，在整体上，开放数据浏览率与下载率呈弱相关。这表明在 10 大主题中，浏览率与下载率的值都较高的情况鲜有发生，然而，浏览率与下载率同为衡量用户对开放数据利用状况的重要指标，二者呈强正相关且其值同时较高才是数据利用的最佳状态，因此政府应积极采取相关措施提高浏览率和下载率的水平。

②其次，在主题视角下，教育科技、民生服务、经济工商等与社会民生领域相关的主题利用程度较高，其浏览率与下载率呈正相关。这几类数据由于与公众日常生活联系密切，受到的关注度自然更高，在“互联网+”电子政务的时代，政府应继续利用自身优势了解、分析并最大限度满足用户需求；但同时社保就业和能源环境主题浏览率与下载率不相关，其变化趋势呈反向发展，其他大多主题数据呈弱相关。社保就业、能源环境、公共安全等主题同样与用户的生活息息相关，但其利用状况并不理想，这与开放数据集的数量、用户需求等因素有关，因此建议详细分析造成此现象的原因，同时加强开放数据的宣传与引导，有针对性的改善此类数据的利用状况；同时，主题视角下开放数据的整体特征与部分特征具有不一致性，因此在改善开放数据整体利用状况的同时，还应注意其内部数据的增值。

③最后，在区域视角下，济南、上海开放数据的浏览率和下载率呈正相关，其中上海开放数据利用度排名首位，济南则排在末位，其余各地如贵阳，虽然开放数据浏览率和下载率都较高，但二者之间线性关系较弱，我国各城市政府开放数据的利用呈现出不均衡性的特点；此外，其开放数据的整体特征与部分特征大体上具有一致性。开放数据利用的不均衡性受到其发展进程、公众知晓度、利用环境与社会经济发展水平等多因素（影响因素将在另文分析）的影响，因此政府应具备开放的思想和意识，加快数据开放的步伐，提高其社会知晓度，在提高开放数据利用率的同时缩小我国各城市之间的利用差距。

6 不足及展望

本文从微观视角出发，选取了哈尔滨、济南、上海、武汉、广州和贵阳 6 地政府开放的部分数据资源作为研究样本，在归纳分类的基础上从主题和区域视角上计算分析了开放数据的浏览率和下载率等指标，进一步做聚类分析与相关性分析，发现在不同主题下和不同城市中开放数据的利用现状都呈现出不均衡性，最后提出相关建议。虽然对开放数据的利用状况的衡量与评价尚未有统一标准，仅通过浏览率与下载率反映其利用现状具有一定的局限性，但与纯定性分析或宏观分析的研究相比，本研究以开放数据本身作为切入点，通过实时数据的抓取与定量指标的计算，在一定程度上增强了研究的说服力和可信度，在整体上能为促进开放数据的利用提供较大参考。

同时,本文也存在一些不足。首先,由于未做全样本分析,可能会影响分析的全面性和科学性;其次,仅分析了开放数据的静态利用状况,尚未在动态环境下对其做时间序列分析;最后,未能深入探究影响开放数据利用现状的因素。针对以上不足,将在后续研究中做进一步探讨。总体上,本研究从两个视角分别通过浏览率、下载率等指标参数的计算,及聚类和相关性分析,对开放数据利用状态揭示具有一定的参考意义。

参考文献:

[1] Open government data [EB/OL]. [2017 – 12 – 13]. https://opengovernmentdata. org/.

[2] The open data barometer [EB/OL]. [2017 – 12 – 14]. http://opendatabarometer. org/4thedition/report/.

[3] PARYCEK P, CHTL J, GINNER M. Open government data implementation evaluation [J]. Journal of theoretical & applied electronic commerce research, 2014, 9(2):80 – 99.

[4] MARTIN C. Barriers to the open government data agenda: taking a multi - level perspective [J]. Policy & internet, 2015, 6(3): 217 – 240.

[5] ROSEIRA C. Exploring the barriers in the commercial use of open government data [C]// International conference on theory and practice of electronic governance. The 9th international conference. New York:ACM, 2016:211 – 214.

[6] ALEXOPOULOS C, LOUKIS E, CHARALABIDIS Y. A methodology for determining the value generation mechanism and the improvement priorities of open government data systems [J]. Computer science & information systems, 2016, 13(1):237 – 258.

[7] ATTARD J, ORLANDI F, SCERRI S, et al. A systematic review of open government data initiatives [J]. Government information quarterly, 2015, 32(4):399 – 418.

[8] ZELETI F A, OJO A, CURRY E. Exploring the economic value of open government data [J]. Government information quarterly, 2016, 33(3):535 – 551.

[9] JETZEK T, AVITAL M, BJØRNANDERSEN N. Generating value from open government data[C]// Conference on information systems (ICIS 2013): reshaping society through information systems design. Milano:ICIS, 2013.

[10] GRAVES A, HENDLER J. A study on the use of visualizations for open government data [J]. Information polity, 2014, 19(1):73 – 91.

[11] 郑磊,关文雯. 开放政府数据评估框架、指标与方法研究[J]. 图书情报工作,2016,60(18):43 – 55.

[12] 陈水湘. 基于用户利用的政府数据开放平台价值评价研究——以19家地方政府数据开放平台为例[J]. 情报科学,2017,35(10):94 – 98,102.

[13] 武琳,伍诗瑜. 城市开放政府数据平台服务绩效评估体系构建及应用[J]. 图书馆论坛,2018,38(2):59 – 65.

[14] 刘新萍, 肖鑫, 黄奕奕. 中国地方政府环境数据开放的现状、问题与对策:基于国内部分省市开放数据平台的分析[J]. 电子政务, 2017(9):30 – 40.

[15] 曹雨佳. 政府开放数据生存状态:来自我国19个地方政府的调查报告[J]. 图书情报工作,2016,60(14):94 – 101.

[16] 汪雷,邓凌云. 基于大数据视角的政府数据开放保障机制初探[J]. 情报理论与实践,2017,40(2):77 – 79.

[17] 马海群,蒲攀. 国内外开放数据政策研究现状分析及我国研究动向研判[J]. 中国图书馆学报,2015,41(5):76 – 86.

[18] 王法硕,王翔. 我国政府数据开放利用的影响因素与实现路径——一项基于扎根理论的质性研究[J]. 情报杂志,2016,35(7):151 – 157.

[19] 黄如花,王春迎. 英美政府数据开放平台数据管理功能的调查与分析[J]. 图书情报工作,2016,60(19):24 – 30.

[20] 陈美. 政府数据开放利用:内涵、进展与启示[J]. 图书馆建设,2017(9):44 – 50,77

[21] 徐慧娜,郑磊. 面向用户利用的开放政府数据平台:纽约与上海比较研究[J]. 电子政务,2015(7):37 – 45.

[22] 张子良,马海群. 我国政府数据开放平台利用效果比较研究[J]. 数字图书馆论坛,2016(6):8 – 15.

[23] 姚家奕. 数据仓库与数据挖掘技术原理及应用[M]. 北京:电子工业出版社,2009.

[24]《2017中国地方政府数据开放平台报告》[EB/OL]. [2017 – 05 – 28]. http://www. cbdio. com/BigData/2017-05/28/content_5528780. htm.

作者贡献说明:

段尧清:提出研究思路,修改论文;

邱雪婷:数据收集与分析,撰写论文;

何思奇:数据收集与分析。

Analysis on the Status of China's Urban Government Open Data Utilization from the Thematic and Regional Perspectives

Duan Yaoqing Qiu Xueting He Siqi

School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] This paper aims to analyze the use status of urban government open data in China from the perspective of subject and region, and explore the linear relationship between the degree of attention and utilization of open data, so as to improve the efficiency of the use of open data by our government. [**Method/process**] It se-

lects six places such as Harbin, Jinan, Shanghai, Wuhan, Guangzhou and Guiyang as examples to conduct statistical comparative analysis, cluster analysis and regression analysis on multiple indicators such as browsing rate of the government open data to reveal the utilization status of China's urban open data under the thematic and regional perspectives, and to explore the linear relationship between browsing rate and download rate. [**Result/conclusion**] The use of open data by urban governments in China has the following characteristics: The overall open data browsing rate is weakly related to the download rate. From the perspective of the theme, educational technology, people's livelihood services, economic and industrial, and other topics related to the social and people's livelihood are highly utilized, and their browsing rate is positively correlated with the download rate; the overall characteristics of open data are inconsistent with some features. From a regional perspective, the browsing rate and download rate of Jinan and Shanghai are positively correlated. Among them, Shanghai ranks first in the use of open data, while Jinan ranks in the bottom. Guiyang's open data browsing rate and download rate are both high, but both are Weak correlation; the overall characteristics of open data and some features are generally consistent.

Keywords: open government data　correlation analysis　utilization